Introduction to Issue & Solution

- Start ups play a major role in economic growth which help to bring in new ideas, spur innovation, create employment and better the lives of society through their innovation.

- With an exponential growth in start ups over the past few years, it is increasingly important to build a business success prediction model to predict the success of a start ups so that financial investments in such start ups would not be in vain.

- This would allow investors or venture capitalists to source for companies that have the potential for rapid growth, thereby allowing them to make a more informed decision on their investments and be a step ahead of the competition.

- The objective of this project is to predict whether a start up which is currently operating turns into a success or a failure.

- The success of a company is defined as the event that gives the company's founders a large sum of money through the process of M&A (Merger and Acquisition) or an IPO (Initial Public Offering).

- This project achieves to find similar details that encompasses a successful start up through machine learning.

Information on the dataset

- The dataset includes information about the companies and investment funds, such as the the type of industry the company is in, the number of funding rounds, amount of collected funds, and investment type (seed, angel funding, series A, B, C, etc.).

Exploring the Data
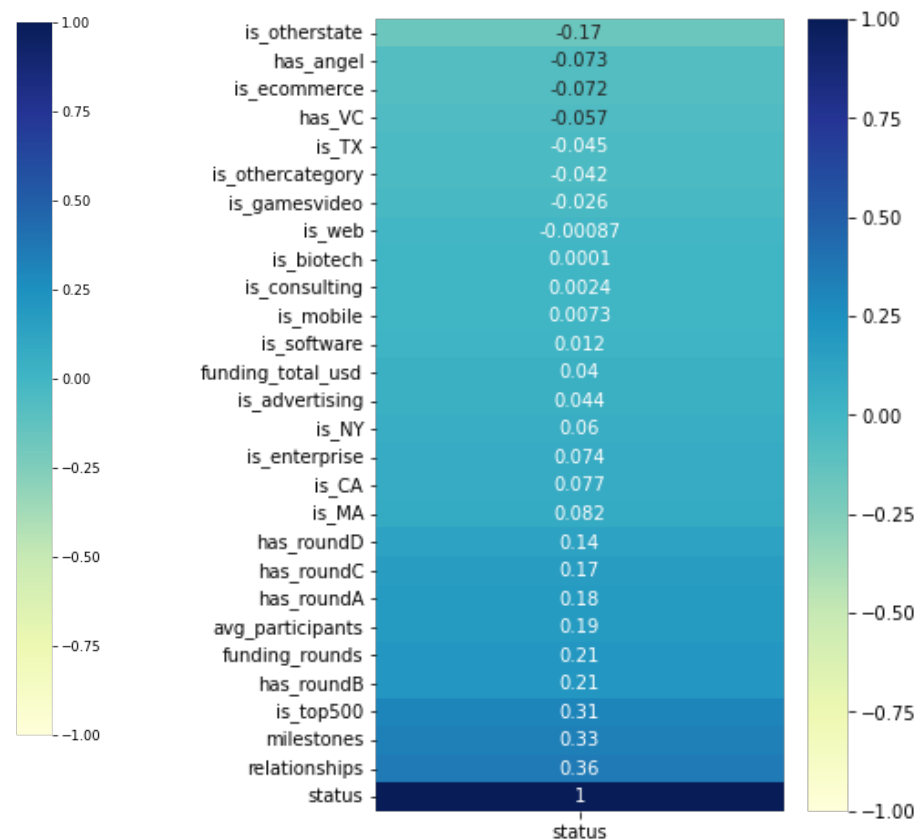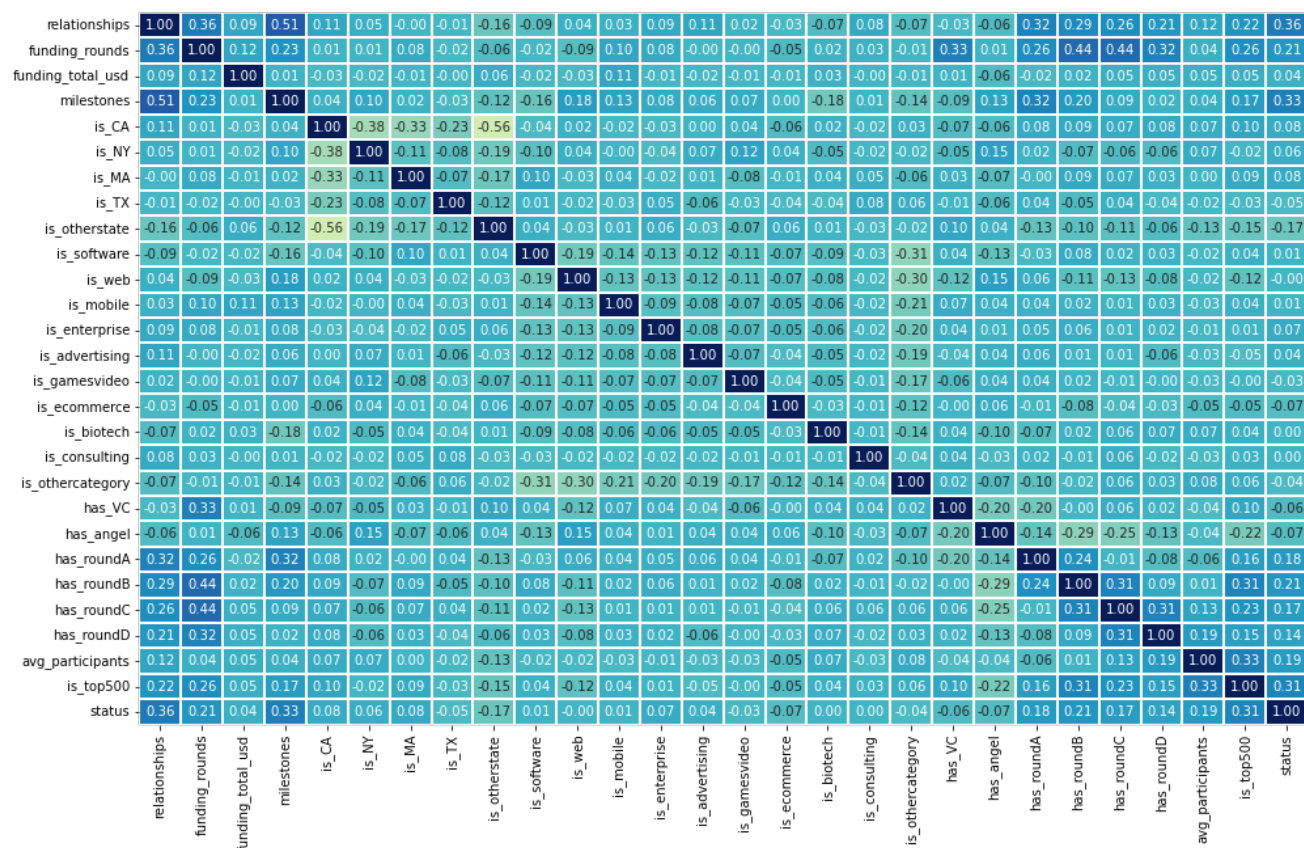
Not much correlation is present in between many of the features available. The highest positive score of **0.51** between relationships and milestones suggests a small correlation in both features.

Checking correlation with 'Status' column - Negative correlation: 8 features; Positive correlation: 19 features.
It seems like relationships, milestones, achievements of a start ups have higher correlation to a successful status than other features.
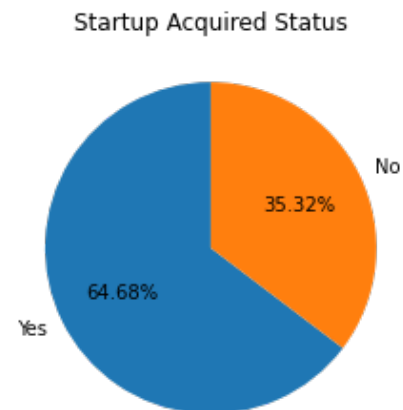
Exploring the Data
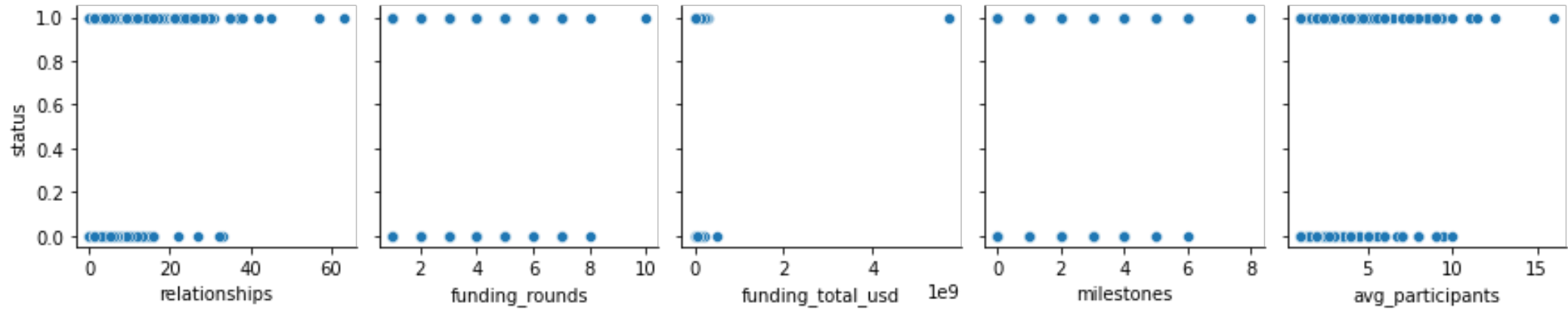
From the above, it is more apparent that successful start ups usually have a greater number of relationships and average participants. It may also potentially suggest that start ups with more funding rounds, milestones and total funding have better prospects as well.
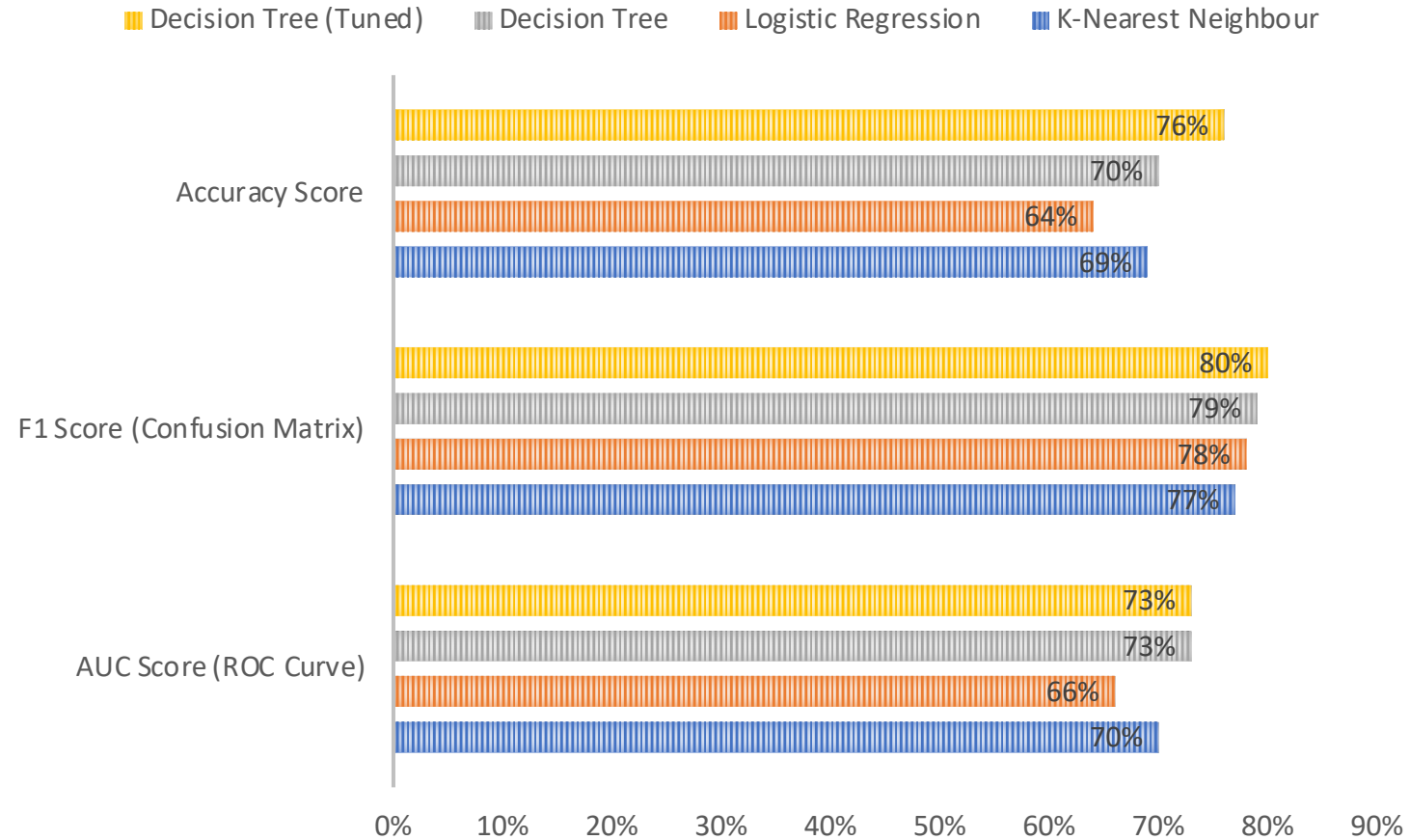
Percentage of successful and unsuccessful startups within dataset

Classification Models

CLASSIFICATION MODEL COMPARISON

Legend: Decision Tree (Tuned), Decision Tree, Logistic Regression, K-Nearest Neighbour

Accuracy Score
- Decision Tree (Tuned): 76%
- Decision Tree: 70%
- Logistic Regression: 64%
- K-Nearest Neighbour: 69%

F1 Score (Confusion Matrix)
- Decision Tree (Tuned): 80%
- Decision Tree: 79%
- Logistic Regression: 78%
- K-Nearest Neighbour: 77%

AUC Score (ROC Curve)
- Decision Tree (Tuned): 73%
- Decision Tree: 73%
- Logistic Regression: 66%
- K-Nearest Neighbour: 70%

Conclusion

In Summary

- Utilized Decision Tree, Logistic Regression and K-Nearest Neighbour models to try to predict the success of start ups.
- Decision Tree Model has achieved the highest accuracy, F1 score and AUC score.
- With GridSearch, we are able to improve on parameters of the Decision Tree Model which can lead to higher scores.

Limitations in Analysis

- Dataset is a relatively new and small dataset, thus results might be skewed and the dataset might contain irrelevant features.
- Model can be further improved by experimenting more with the hyperparameters in the model for better accuracy.
- Consideration factors should be strongly correlated in order to enhance prediction accuracy
- Decision tree could have been improved in more aspects such as pruning to minimize overfitting
- Analysis using more robust models such as Random Forest could have also been used to reflect more accurate and relevant predictions.